МІНІСТЕРСТВО ОСВІТИ І НАУКИ, МОЛОДІ ТА СПОРТУ УКРАЇНИ

НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ

«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ»

ФІЗИКО-ТЕХНІЧНИЙ ІНСТИТУТ

Комп’ютерний практикум №1

з дисципліни Криптографія:

«Експериментальна оцінка ентропії на символ джерела відкритого тексту»

Виконали:

Студенти групи ФБ-96

Сендецький Костянтин

Твердохлібов Денис

Київ 2021

Мета роботи

Засвоєння понять ентропії на символ джерела та його надлишковості, вивчення та порівняння різних моделей джерела відкритого тексту для наближеного визначення ентропії, набуття практичних навичок щодо оцінки ентропії на символ джерела.

Завдання

1. Написати програми для підрахунку частот букв і частот біграм в тексті, а також підрахунку H1 та H2 за безпосереднім означенням. Підрахувати частоти букв та біграм, а також значення H1 та H2 на довільно обраному тексті російською мовою достатньої довжини (щонайменше 1Мб), де імовірності замінити відповідними частотами. Також одержати значення H1 та H2 на тому ж тексті, в якому вилучено всі пробіли.

2. За допомогою програми CoolPinkProgram оцінити значення (10) H , (20) H , (30) H .

3. Використовуючи отримані значення ентропії, оцінити надлишковість російської мови в різних моделях джерела.

Хід роботи

Завдання 1

В цій частині було визначено частоти простих(з кроком 2) і перехресних біграм(з кроком 1). За цими показниками була визначена ентропія для біграм і літер. Частота біграм отсортована по їх буквенним значенням а для літер за зростанням частоти

Частоти літер

З пробілами	Без пробілів

	З пробілами
щ	0.167266
т	0.0958129
е	0.0729672
а	0.0690919
х	0.0578961
д	0.0553795
г	0.0499397
ь	0.0443241
м	0.0418346
	0.039269
л	0.0332092
с	0.0285882
р	0.0245607
ы	0.02391
ч	0.0224936
в	0.0201132
б	0.0179647
э	0.0166804
у	0.0154645
к	0.0151572
н	0.014582
ф	0.0140794
о	0.0136399
и	0.00945074
з	0.0087676
п	0.00711991
ш	0.00648101
ю	0.0051985
ц	0.00296088
ж	0.00239159
й	0.00235561
я	0.00105067

	Без пробілів
в	0.115058
и	0.0876236
а	0.0829699
т	0.0695253
з	0.0665032
ч	0.0599707
г	0.0532272
ы	0.0502376
к	0.0471566
ь	0.0398797
м	0.0343305
е	0.0294941
х	0.0287127
я	0.0270117
л	0.0241532
щ	0.0215731
ц	0.0200309
э	0.0185708
р	0.0182017
н	0.017511
п	0.0169074
д	0.0163796
о	0.011349
с	0.0105287
ж	0.00855004
й	0.00778281
б	0.00624269
ю	0.00355561
у	0.00287198
ф	0.00282876
ш	0.00126171

біграми без пробілів(20 найчастіших)

Перехресні

	Перехресні біграми без пробілів
то	0.0174345
на	0.0134063
ст	0.0128056
он	0.0126965
но	0.011925
ал	0.0119179
не	0.0118499
ов	0.0113221
ен	0.0104954
го	0.0100108
ко	0.00975863
по	0.00970975
ни	0.00958011
ос	0.00953406
от	0.0093272
ка	0.00926485
ла	0.00918905
во	0.00879587
ро	0.00827518
ра	0.008193

Прості

	Прості біграми без пробілів
то	0.0174387
на	0.0134772
ст	0.0127999
он	0.0126511
ал	0.0119654
не	0.011835
но	0.0117458
ов	0.0113547
ен	0.0104238
го	0.0101022
по	0.00982309
ко	0.00978342
ни	0.00956239
ос	0.00947029
ка	0.00932719
от	0.00926343
ла	0.00913025
во	0.00873636
ро	0.00837931
ра	0.00814695

Біграми з пробілами(20 найчастіших)

Перехресні

	Перехресні біграми з пробілами
о	0.0239
е	0.0188396
а	0.0186661
и	0.0180774
н	0.0163707
с	0.016117
в	0.0147737
то	0.0142505
п	0.0140404
о	0.0134594
я	0.0111085
и	0.0110495
на	0.0108666
ст	0.0104188
ь	0.0104064
не	0.0097817
но	0.00962832
ал	0.00951387
к	0.00903543
т	0.00829566

Прості

	Прості біграми з пробілами
о	0.0239596
е	0.018787
а	0.0186207
и	0.0179352
н	0.0162409
с	0.0162409
в	0.0147802
то	0.0143295
п	0.0140782
о	0.0134328
я	0.0111887
и	0.0110813
на	0.0109386
ь	0.0104359
ст	0.0103663
не	0.00980235
но	0.00957463
ал	0.0095416
к	0.00897644
го	0.00829094

Ентропія для букв без пробілів: 4.4392968078807

Надлишковість: 0.10393199426244926

Ентропія для букв з пробілами: 4.348166097136004

Redundancy: 0.13036678057279916

Ентропія для перехресних біграм без пробілів: 4.122027697282843

Надлишковість: 0.16797247443734187

Ентропія для простих біграм без пробілів: 4.12094557882284

Надлишковість: 0.168190899060067

Ентропія для перехресних біграм з пробілами: 3.9422918405387293

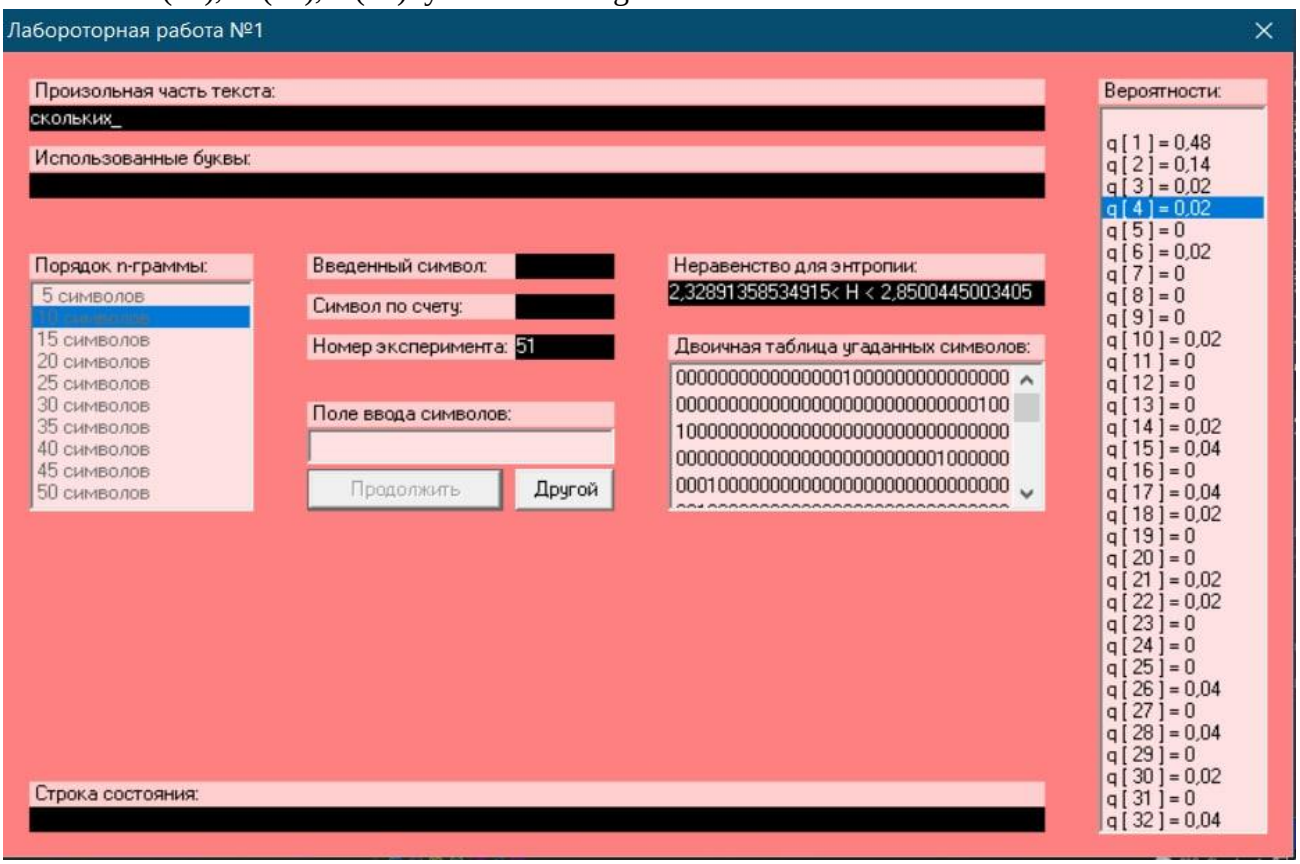
Надлишковість: 0.21154163189225417

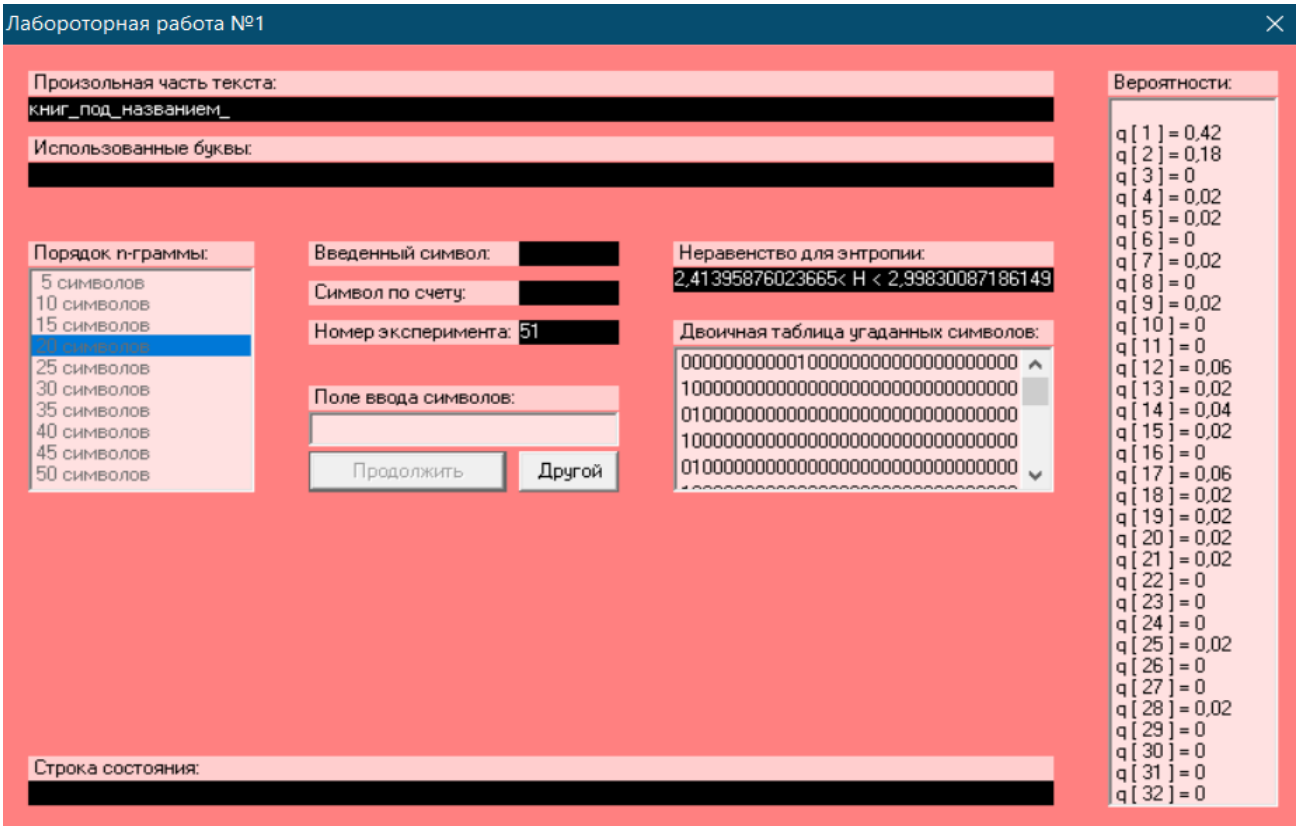
Ентропія для біграм з пробілами: 3.9419919173607862

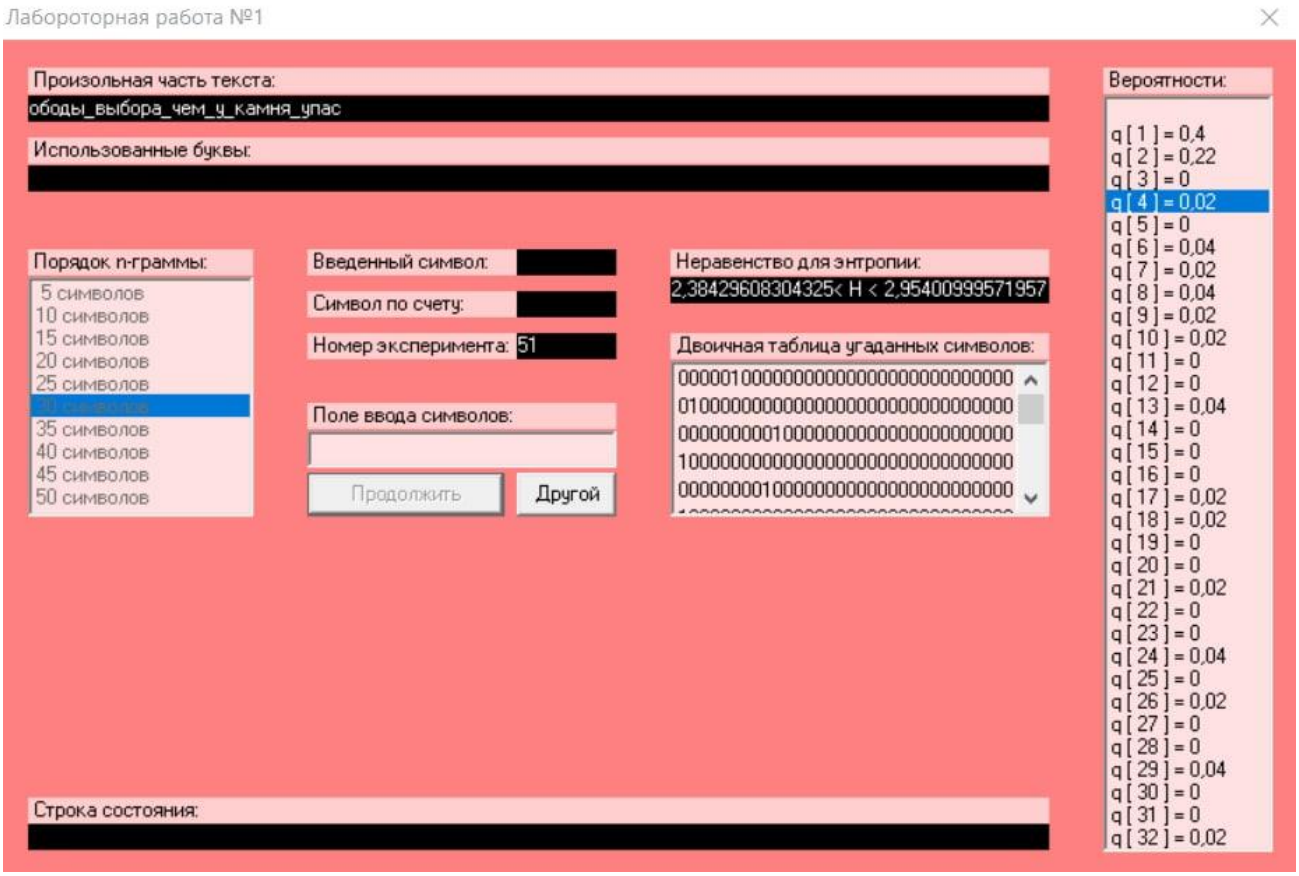
Надлишковість: 0.2116016165278427

Завдання 2

Значення H(10), H(20), H(30) у CoolPinkProgram







Отримали значення

2.3289 < H(10) < 2.8500

2.4139 < H(20) < 2.9993

2.3842 < H(30) < 2.9540

Знаходимо значення надлишковості джерела відкритого тексту

0,5342 > R(10) > 0,4300

0,5172 > R(20) > 0,4001

0,5231 > R(30) > 0,4092

Висновки: Засвоїли поняття ентропії на символ джерела та його надлишковості, вивчили та порівняли різниі моделі джерела відкритого тексту для наближеного визначення ентропії, набули практичних навичок щодо оцінки ентропії на символ джерела.